Markus I. Eronen, M.A.

Replacing Functional Reduction with Mechanistic Explanation

Markus I. Eronen
Institut für Kognitionswissenschaft
Universität Osnabrück
49069 Osnabrück
Tel.: 0176 2618 6314
E-mail: maeronen@uos.de

Abstract

Recently the functional model of reduction has become something like the standard model of reduction in philosophy of mind. In this paper, I argue that the functional model fails as an account of reduction due to problems related to three key concepts: functionalization, realization and causation. I further argue that if we try to revise the model in order to make it more coherent and scientifically plausible, the result is merely a simplified version of what in philosophy of science is known as mechanistic explanation. Hence, instead of analyzing reduction in philosophy of mind in terms of functional reduction, it should be analyzed in terms of mechanistic explanation.

1. Introduction

In recent years, the functional model of reduction has become something like a standard model of reduction in philosophy of mind. However, the model is by no means new: its main ideas are already present in the filler-functionalism of David Lewis (1972). Lewis' idea was roughly that a given mental state *M* is defined functionally in terms of its causal role, but in the end *M* is nothing more than the physical states that occupy this role. For instance, "pain" is a functional concept specified by its causal role, but in the end pain *just is* the physical (neural) state that fills that causal role. This physical state can be one thing in humans, another in octopuses, and still something else in Martians. These different states are all picked out by the functional concept "pain", which (non-rigidly) designates different physical fillers in different species.

More recently philosophers like Joseph Levine (1993), David Chalmers (1996), Frank Jackson (Chalmers and Jackson 2001), and Jaegwon Kim (1998, 2005) have presented somewhat varying models of functional reduction based on this general approach. All of these authors have then applied the supposedly general model of reduction to the problem of phenomenal consciousness, arguing that phenomenal properties are fundamentally irreducible, or that there is an "explanatory gap" between phenomenal properties and the physical domain.

I will focus here on Kim's version of functional reduction, since it is exceptional in its clarity, and has also been extremely influential. I will argue that the functional model fails to capture the nature of reduction in psychology and neuroscience. Furthermore, I will show that if we try to revise the functional model in order to make it more scientifically credible, it turns out that the revised model is not significantly different from mechanistic explanation. Hence, instead of analyzing reduction in philosophy of mind in terms of functional reduction, it should be analyzed in terms of mechanistic explanation.

2. The functional model

Kim's main motivation for invoking the model of functional reduction is to show that mental properties (with the exception of phenomenal properties) can be saved from the *causal exclusion argument*, which I will briefly sketch here. Several different versions of the argument exist; the formulation here reflects Kim's most recent accounts (Kim, 2002, 2005). The argument is based on certain principles that together create a problem for mental causation (Kim, 2002, 278):

> *The Problem of Mental Causation*: Causal efficacy of mental properties is inconsistent with the joint acceptance of the following four claims: (1) physical causal closure, (2) exclusion, (3) mind-body supervenience, and (4) mental/physical property dualism (i.e., irreducibility of mental properties).

The principle of physical causal closure states that every physical occurrence has a sufficient physical cause. The principle of exclusion states that no effect has more than one sufficient cause, except in cases of genuine overdetermination, such as two bullets hitting the heart of a victim at exactly the same time, both causing death.

It is easy to see how the four principles taken together lead to trouble. Let us start by assuming that (the instantiation of) a mental property *M* causes (the instantiation of) another mental property *M**. Due to mind-body supervenience, *M* supervenes on some physical property *P*, and *M** supervenes on some physical property *P**. Since *M** supervenes on *P**, *M** must be necessarily instantiated whenever *P** is instantiated, no matter what happened before: the instantiation of *P** alone necessitates the occurrence of *M**. Thus, according to Kim, the only way that *M* can cause *M** is by causing *P**.

This is where the principle of causal closure kicks in: *P** must also have a sufficient physical cause. This means that *P** has a sufficient physical cause *P* and a mental cause *M*, and the exclusion principle states that one of these must go – if we would accept cases like this as genuine overdetermination, we would get massive overdetermination of physical effects by mental causes, which is highly implausible. Obviously *M* is the one that has to go, since if *M* was the only cause of *P**, this would violate the principle of physical causal closure. Therefore, *M* cannot be the cause of *M** or of any other mental or physical property. This holds for all mental properties, and we

have the striking conclusion that, under mind-body supervenience, mental properties are causally impotent.

According to Kim, physical causal closure and mind-body supervenience are among the inescapable commitments of all physicalists. The exclusion principle is taken to be a general metaphysical constraint that can hardly be challenged. This leaves only mental/physical property dualism (i.e., the irreducibility of mental properties) as the principle that has to go. Therefore, Kim's conclusion is what he calls "conditional reductionism": "If mentality is to have a causal influence in the physical domain – in fact, if it is to have any causal efficacy at all – it must be physically reducible" (Kim, 2005, 161).

What does reduction then amount to? Kim's answer is the functional model:

> To reduce a property, say being a gene, on this model, we must first "functionalize" it; that is, we must define, or redefine, it in terms of the causal task the property is to perform. Thus, being a gene may be defined as being a mechanism that encodes and transmits genetic information. That is the first step. Next, we must find the "realizers" of the functionally defined property – that is, properties in the reduction base domain that perform the specified causal task. It turns out that DNA molecules are the mechanisms that perform the task of coding and transmitting genetic information – at least, in terrestrial organisms. Third, we must have an explanatory theory that explains just how the realizers of the property being reduced manage to perform the causal task. In the case of the gene and the DNA molecules, presumably molecular biology is in charge of providing the desired explanations. (Kim, 2005, 101)

Kim presents the functional model as a better and more scientifically credible alternative to Ernest Nagel's (1961) classic but problematic model: "Nagel reduction of pain requires an all-or-nothing, one-shot reduction of pain across all organisms, species, and systems. It is clear that functional reduction gives us a more realistic picture of reduction in the sciences" (Kim, 2005, 102). In Nagel's model, reduction of a theory T_2 consists in deducing it from a more fundamental theory T_1, with the help "bridge laws" that connect the terms of the two theories. What Kim sees as the main problem with Nagel's model is that it gives us reductions that do not explain (Kim, 1998, 90-97, 2005, 98-101). This is

because, according to Kim, the reductive work in Nagel's model is done by the biconditional bridge laws that connect properties of the reduced theory to properties of the reducing theory, and these bridge laws are just "unexplained auxiliary premises" that are themselves in need of explanation.

Ausonio Marras (2002) has pointed out that bridge principles do not in fact play a key role in Nagelian reductions, and therefore Kim's critique is largely misplaced. However, in the present context, Nagelian reduction faces other, more fundamental, problems. The main problem of Nagelian models of reduction in the context of psychology and neuroscience is that they require the theories involved in reductions to be formalized, either according to the syntactic (e.g, Nagel, 1961) or the structuralist/semantic (e.g., Bickle, 1998) view of theories.[1] The problem is that while formal theories that are suitable as starting points of logical derivations may be available in theoretical physics, most special sciences simply do not have any well-structured theories that could be handled formally. Rather than trying to formulate such theories, psychologists and neuroscientists typically look for descriptions of mechanisms that can serve as explanations for patterns, effects, capacities, phenomena, etc., and this explanatory enterprise at best involves fragments of formal theories (Craver, 2007, Machamer et al., 2000, McCauley, 2007, Walter and Eronen, forthcoming). Furthermore, some generally accepted cases of scientific reduction – for instance the reduction of genetics to molecular biology — do not seem to involve formal theories (Sarkar, 1992). In this light, the model of functional reduction is prima facie promising, since it is a model of property reduction, not theory reduction, and does not require formal theories.[2]

Let us then take a closer look at Kim's model of functional reduction (Kim, 1998, 97-103, 1999, 10-13). The reduction of property *M* consists of three steps:

Step 1: *M* must be functionalized – that is, *M* must be construed, or reconstrued, as a property defined by its causal/nomic relations to other properties. As Kim puts it:

> [W]e must first "prime" *M* for reduction by construing, or reconstruing, it *relationally* or *extrinsically*. This turns *M* into a relational/extrinsic property. For functional reduction we construe *M* as a second-order property defined by its causal role – that is, by a causal

> specification *H* describing its (typical) causes and effects. So *M* is now the property of having a property with such-and-such causal potential[.] (Kim, 1998, 98)

Thus, property *M* is defined as a "second-order" property: it is a property that some first-order properties have.

Step 2 consists of finding the realizers of *M*. These are the first-order properties in the reduction base domain that have the right causal/nomic relations, i.e., the properties that fit the causal specification *H*. The realizers can be different in different systems, allowing for multiple realizability. Step 2 is a matter of scientific research, or as Kim puts it, "a scientifically significant part of the reductive procedure" (Kim, 1999, 11).

Step 3 is to find a theory that explains how the realizers actually perform the causal role specified in Step 1. Like Step 2, Step 3 is also a matter of scientific research, and these steps are intertwined, since figuring out what the realizers of *M* are certainly involves theories about the causal/nomic relations in the reduction base.

One of the central points of Kim's account is that functionally reduced properties are nothing "over and above" the reducing properties: "Central to the concept of reduction evidently is the idea that what has been reduced need not be countenanced as an independent existent beyond the entities in the reduction base – that if *X* has been reduced to *Y*, *X* is not something 'over and above' *Y*" (Kim, 1999, 15). According to Kim, this means that reduction has to lead either to identities (conservative reduction) or eliminations (replacement / eliminative reduction). Is functional reduction then conservative or eliminative?

First of all, Kim argues that when *M* has been functionally reduced to *P*, *instances* of *M* can be identified with the instances of *P* (Kim, 1999, 15-16). He invokes the "causal inheritance principle", which states that "[i]f a functional property [*M*] is instantiated on a given occasion in virtue of one of its realizers, [*P*], being instantiated, then the causal powers of this instance of [*M*] are identical with the causal powers of this instance of [*P*]." If we accept this principle, it follows that the instances of *M* and *P* have exactly the same causal powers, and it is hard *not* to identify the instances, since if they were not identical, the difference could not even be detected. However, what is at issue in the exclusion argument is not token causation (one instance or event causing another instance

or event), but type causation. The problem is whether mental *properties* can have causal powers – in other words, whether some event can cause a mental or physical event in virtue of being an instantiation of a mental property. Therefore, for avoiding the exclusion argument it is not enough that *instances* of *M* are identical to instances of physical properties, also the property *M* itself has to be identical to a physical property *P*.

The situation is made even more complicated if (as is generally assumed) *M* can have multiple realizers. Therefore, Kim sees only two options: we can (1) identify *M* with the disjunction of its realizers, or (2) give up *M* as a real property and only recognize it as a property designator that picks out many different properties (the realizers of *M*).

Identifying *M* with the disjunction of its realizers is problematic. The realizers must have different causal roles, since otherwise they wouldn't be different realizers (Kim supports a causal theory of properties). If *M* is identical to a set of causally and nomologically heterogeneous properties, Kim reasons, then *M* itself must be causally and nomologically heterogeneous, and is unfit to figure in laws, and is thereby not a scientific property (see Kim (1992) for more details of this argument).[3]

Therefore, Kim is inclined to accept the second option:

> One could argue that by forming "second-order" functional expressions by existentially quantifying over "first-order" properties, we cannot be generating new properties (possibly with new causal powers), but only new ways of indifferently picking out, or grouping, first-order properties, in terms of causal specifications that are of interest to us. (Kim, 1999, 17)

This makes functional reduction eliminative: we have to accept that mental properties are not genuine properties in their own right. Kim accepts this only because the other alternatives (disjunctive identities or property dualism) are wrought with major philosophical problems (Kim, 2008, 112). I will return the problems of this option in Section 3.2 below.

3. What Is Wrong with the Functional Model

The functional model has been recently criticized from different angles. Ausonio Marras (2002, 2005) has argued that when we analyze the model carefully and accept certain plausible background assumptions, it in fact leads back to Nagel reduction, which it was supposed to replace. In the same vein but with different arguments, Max Kistler (2005) has argued that functional reduction requires local bridge laws that are left just as unexplained as in a Nagel reduction. John Bickle (2008, forthcoming) does not criticize the model itself, but points out that it is based almost entirely on logical and metaphysical considerations, and that the examples given to support it reflect an elementary school understanding of science. In this sense, the functional model is a step *backward* from Nagelian models, which were at least based on science (though not psychology and neuroscience).

I will develop the last line of argument in more detail, and show that from the point of view of philosophy of science and scientific practice, the functional reduction approach is unacceptable. I will focus on three salient problems of the model. 1) Where do the functional definitions of properties to be reduced come from? (2) What is the "realization" relation between the property to be reduced and the reducing properties? (3) What notion of causation does the model require? These are by no means the only problems or points that need clarification, but they suffice to show why the model fails as a general account of reduction.

3.1. Functionalization

As we have seen, Step 1 in the functional model consists in defining or redefining the property to be reduced in terms of its causal role. However, it is not clear how we get the causal definition of the property to be reduced. Kim seems sympathetic to the view of Chalmers and Jackson (2001) and Levine (1993), according to which reductive explanation requires analytic definitions grounded in (a priori) conceptual analysis (see Kim, 2005, Ch. 4). The first step of functional reduction would thus consist in finding the analytic definition for the property to be reduced through conceptual analysis.

However, if the functional definition of the mental properties is to be based on conceptual analysis that is (at least relatively) a priori, this leads to a fundamental problem: our a priori ideas about our own psychological states or processes are often simply wrong. Consider for example memory. An armchair conceptual analysis would indicate that memory is some kind of a simple storage, where our past experiences are waiting for retrieval – Plato compared memory to an aviary of birds, from which we take the correct bird when memory retrieval is successful, and the wrong bird when it is not. However, scientific research has revealed that memories are not just retrieved, but actively constructed, and subjectively compelling memories sometimes turn out to be radically inaccurate. Furthermore, memory comprises several subsystems (short term memory, long term memory, episodic memory, visual memory, etc.), which neither individually nor taken together correspond to the simple storage envisioned by *a priori* analysis (see Bechtel (2008, Ch. 2) for a detailed philosophical analysis of memory research). Similar considerations apply to pain (Hardcastle, 2001), which has for decades been a standard example in philosophy of mind.

It is thus clear that mere conceptual analysis is not sufficient for working the properties "into shape" for reduction. One has to either allow for scientific revision of common sense definitions of mental properties, or simply focus on properties as defined by empirical psychology.[4]

Furthermore, in both cases we have to allow for the revision and adjustment of the definitions as science proceeds. Such revision and interplay across levels is commonplace in science. One of the first philosophers to emphasize the importance of this co-evolution of concepts and theories was William Wimsatt, drawing from scientific practice in biology:

> A lower-level model is advanced to explain an upper-level phenomenon which it doesn't fit exactly. This leads to a closer look at the phenomenon, and perhaps results in some change in the way in or detail with which it is described. This will also lead to changes in the lower level model and may suggest new phenomena to look for. (Wimsatt, 1976, 231)

Also Bechtel and Richardson (1993) have described in detail the complexities involved in characterizing the phenomena to be explained in biology, based on detailed analyses of cases from history of biology, and one of their points is that scientists often have to constantly redefine the phenomena they are trying to explain. More broadly speaking, in the mechanistic explanation paradigm (Bechtel and Richardson, 1993, Machamer et al., 2000, Craver, 2007, Bechtel, 2008), a crucial point is that there is constant interplay between different levels of explanation, and both top-down and bottom-up influences.

There is also a further problem related to functionalization, even if take empirical psychological properties to be the targets of reduction and allow for constant revision of their functional definitions. It is quite possible that in the end we are unable to find any neuroscientific properties playing the causal role of some psychological properties, and thus we cannot functionally reduce them. The easiest solution in these cases would be to revise the functional definitions of the psychological properties, but this is not always justifiable. We might want to retain some psychological properties more or less as they are, since they are useful in scientific explanations. For example, Khalidi (2005) takes up the psychological property of fear, and shows (based on empirical results in cognitive neuroscience) that distinctions made at the neurophysiological level cross-cut the distinctions made at the psychological level. That is, from the vantage of neurophysiology, there is nothing playing the functional role associated with the psychological state of fear. Importantly, this is not a case of multiple realizability, which is a one-to-many relationship. In this case, there is simply just mismatch: a "one-to-none" relationship. However, we would not want to eliminate or revise the psychological concept of fear, since it still plays an important role in research and scientific explanations.

In this case, it seems that there are no neurophysiological states playing the causal role of fear, and the option of redefining fear does not seem very fruitful. Hence, Step 2 in functional reduction of fear fails. But should we conclude from this that fear is fundamentally irreducible and threatened by the exclusion argument? Or should we eliminate the property of fear from our ontology? Both options seem implausible. The framework of functional reduction seems unsuitable for dealing with situations like this.

Certainly the basic idea that the properties to be reduced have to specified causally is correct and in accordance with scientific practice. However, functionalization is not just a matter of conceptual analysis, it is not even remotely an a priori matter, and functional definitions can change as research proceeds. Furthermore, in some cases we might not be able to find neural realizers that play the functional role definitive of a mental property. This does not mean that Kim's functional model is fundamentally wrong, but it surely is too crude and simplified.

3.2. Realization

The second step in Kim's account of functional reduction is finding the "realizers" of the functionally defined property to be reduced. But what makes some property a realizer of another property? How should we understand this realization relation? And what sorts of things are the realizers of mental properties?

The roots of talk of "realization" in philosophy of mind go back to multiple realizability. Hilary Putnam (1967) famously argued that it is extremely plausible that a given mental state (like "being in pain") can be realized by different physical-chemical states in different organisms. In the debate that followed, very little attention was paid to the notion of realization itself. However, as several philosophers have recently shown (e.g., Polger (2004, 2007), Shapiro (2004)), the realization relation is much more problematic than has been generally assumed. For example, a computer realizing an abstract algorithm or computation can hardly involve the same realization relation as a brain realizing a mental state, since mental states are thought to be individuated causally, but abstract algorithms or computations are not individuated causally (Polger, 2004, 2007). This means that there might be no general realization relation that applies to all the different cases that are presented as typical cases of realization. However, I will not pursue this line of argumentation here, since others (i.e., Polger and Shapiro) have already elaborated it in detail. It might also be that Kim's account does not need any general notion of realization, and that a more "local" notion would suffice. In this section I will show that even if we limit the discussion to psychological properties and their

realizers, and accept that there is no general notion of realization, Kim's notion of realization leads to problems.

Let us consider the case of mental properties and their neural realizers. The mental properties are to be functionally defined in terms of their causal relations to other mental properties. What is it then for a neural property to realize a mental property? According to Kim, the realizers have to perform the causal task specified in Step 1, that is, they have to "occupy" or "fill" or "play" the causal role definitive of the mental property.

But what does this mean? If we take the realizers to be properties, it seems that the only way to make sense of this is that the realizing neural property has to be embedded in a causal structure that is isomorphic to the causal structure in which the mental property is embedded. That is, the causal "context" of the neural property has to be isomorphic to the causal "context" of the mental property. What else could it mean for the neural property to occupy the causal role definitive of the mental property?

However, this leads to problems, since Kim's aim is to reduce all (non-phenomenal) psychological properties, not just one of them. This implies that, in order to accomplish a psychoneural reduction, we would have to figure out the causal structure of all the mental properties we want to reduce, and then find an isomorphic causal structure among the neural properties. If we also assume that laws underlie causal relations, and that theories are sets of laws (both assumptions are controversial, but commonly accepted in philosophy of mind), the implication is that Kim's model comes very close to theory reduction: in order to reduce a psychological theory, we need to find in (or derive from) the neuroscientic theory a structure that is isomorphic to the psychological theory. This is not so different from the "New Wave" model of psychoneural reduction (e.g., Hooker, 1981, Bickle, 1998), where a psychological theory is reduced by deriving from neuroscience an "analogue" or "equipotent image" that is isomorphic to the psychological theory.

Marras (2002, 2005) makes a similar point with a somewhat different reasoning: in a closer analysis, Kim's model turns out to be a model of intertheoretic reduction. If this is the case, the functional model only appears to be an advance over the intertheoretic models, and faces exactly the same problems (see section 2 above).

Another fundamental problem with Kim's notion of realization was already briefly mentioned at the end of section 2: if we accept multiple realizability, the realized properties have to be either identical to the disjunction of the realizers, or just concepts (or predicates or designators). Kim rejects the first option for philosophical reasons and accepts the second one. However, in the context of realization, the problem with second option is that it seems to leave no room for the idea that neural properties realize mental properties. According to the second option, the mental concepts simply (non-rigidly) designate different neural properties in different species, just like in Lewis' (1972) filler-functionalism. If this is true, there is no realization relation here. Mental properties cannot be realized, since *there are no* mental properties, just mental concepts (or property designators) that group physical properties in interesting ways.[5] And mental concepts cannot be realized, since concepts in general are not the sorts of things that are realized. But if this is the case, the whole talk of realization has been misleading, and the claim that the functional model can accommodate multiple realizability turns out false.[6]

Perhaps, however, there are yet other ways of understanding realization. As Polger and Shapiro (2008) have pointed out, one problematic assumption that underlies many of these issues is the assumption that the realizers have to be *properties*. Particularly in more recent writings, Kim himself has been less strict and allows the realizers to be *mechanisms*: "Find the properties (*or mechanisms*) in the reduction base that perform the causal task *C*" (Kim, 2005, 102, my emphasis).

If we (unlike Kim) take this idea of mechanistic realization seriously, it leads to a more complicated picture of mental realization than the one the functional model presents. The idea is that a functionally (causally) defined psychological state, property, or capacity is realized by a neural mechanism that plays that functional role.[7] A crucial aspect of this kind of mechanistic realization is the *multilevel* nature of the mechanisms: on any reasonable understanding of neural mechanisms, they have to be hierarchically organized into levels. Therefore, instead of a simple two-level model with the mental property and its neural realizers, we have a more complicated picture where the realizer is also organized into levels.

An often-cited example of a psychological property or capacity that is realized by a multilevel neural mechanism is memory consolidation (Craver, 2007). Memory

consolidation can be functionally defined in psychological terms as the transformation of short-term memories into long-term memories. A key component in the neural mechanism realizing it is Long Term Potentiation (LTP), a well-studied cellular and molecular phenomenon that exhibits features that are closely connected to memory consolidation. Craver (2007, 165-170) defines the following four levels in the case of spatial memory and LTP: the level of spatial memory, the level of spatial map formation, the cellular-electrophysiological level, and the molecular level. They are levels of composition, where the relata are behaving mechanisms at higher levels and their components at lower levels. Levels of mechanisms in general are local and case-specific, and not intended as universal divisions of nature or science.

In fact, instead of making a distinction between realized properties and realizer properties, it is more appropriate to consider psychological properties simply as higher-level properties of neural mechanisms. For example, it is quite natural to consider psychological properties of memory consolidation as properties at the highest level of the memory consolidation mechanism. In this sense, they are neither identical to the realizing mechanism nor "just concepts" – they are real higher-level properties. The general idea is (roughly) that psychology defines and discovers functional properties that are then integrated into multilevel mechanistic explanations (and undergo "co-evolution" as science proceeds in different disciplines).

In this account, there is no special metaphysical realization relation. Indeed, no such relation is needed for understanding reductive explanation (see section 4). Whether there is multiple realizability in the sense of one-to-many mappings from higher to lower levels is an issue that has little to do with the possibility of reductive explanation.[8] Talk of realization can be preserved, as long as it is understood in a weak or metaphoric sense: a functionally defined psychological property is "realized" by the underlying neural mechanism in the sense that the activity of the mechanism constitutes the function that is definitive of the psychological property.

The key requirement for "realization" in the functional model is that it would somehow save mental causation. In Kim's account, the only way this could work is that the "realized" properties turn out to be just concepts. Now if we adopt the mechanistic approach to realization, and see the realizers as multilevel mechanisms, what happens to

mental causation? Aren't the multilevel mechanisms problematic regarding causation? In the next section, I will argue that the answer is no.

3.3. Causation

As we have seen, the properties to be reduced are defined by their *causal* roles; they are reduced by finding the first-order properties that have that *causal* role; the aim of functional reduction is to save mental properties from the *causal* exclusion argument; reduced properties have no *causal* powers of their own, and so on. Causal notions seem to play a key role in Kim's account. Indeed, the whole motivation for developing the functional model comes from the causal exclusion argument and from worries regarding the causal efficacy of mental properties. But what is causation? What does it mean to say that *X* causes *Y*?

The kind of notion causation Kim has in mind is very strong and robust:

> We care about mental causation, it seems to me, chiefly because we care about human agency, and evidently agency involves a productive/generative notion of causation. An agent is someone who brings about a state of affairs for reasons. If there indeed are no productive causal relations in the world, that would effectively take away agency—and our worries about mental causation along with it. (Kim, 2009, 44)

As the quote indicates, Kim thinks of causation as a relation where the cause generates, produces, or brings about the effect. According to Kim, a weaker account of causation in terms of, for example, counterfactual relations would not be satisfactory, since we would still need the metaphysical account of what makes the counterfactuals we want for mental causation true (Kim, 1998, 71).

In the next section, I will first briefly present one such weaker account of causation, and then argue that, *contra* Kim, this is all we need for understanding mental causation. In the section after that, I will argue that on this account the causal exclusion

argument does not threaten mental causes, and thus a large part of the motivation for the functional reduction of mental properties fades away.

3.3.1. Interventionist Causation

In recent years, several philosophers have presented accounts of causation in terms of interventions and manipulability (Pearl, 2000, Woodward 2003, 2008, Woodward and Hitchcock, 2003, also Spirtes, Glymour, and Scheines, 1993). I will focus here on James Woodward's (2003) version, which is exceptional in its scope and clarity. The guiding insight of the account is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. To put it very roughly, in this model a necessary and sufficient condition for *X* to cause *Y* or to figure in a causal explanation of *Y* is that the value of *Y* would change under some intervention on *X* (in some background circumstances).

An intervention can be thought of as an (ideal or hypothetical) experimental manipulation carried out on some variable *X* (the independent variable) for the purpose of ascertaining whether changes in *X* are causally related to changes in some other variable *Y* (the dependent variable). Of course, several restrictions on interventions must be added – see Woodward (2003) for details. Interventions are not only human activities, there are also "natural" interventions, and the definition of an intervention makes no essential reference to human agency. This sets the interventionist account clearly apart from previous manipulability theories of causation (e.g., Menzies and Price, 1993).

According to Woodward, causal relationships are relationships that are *invariant* under interventions. Suppose that there is a relationship between two variables that is represented by a functional relationship $Y = f(X)$. If the same functional relationship *f* holds under a range of interventions on *X*, then the relationship is invariant within that range. For example, the ideal gas law "$PV = nRT$" continues to hold under various interventions that change the values of the variables (*P*, *V*, and *T*), and is thus invariant within this range of interventions. One consequence of this model is that relata of causation must be represented as variables, but states or properties can easily be

represented as binary variables, such that, for example, 1 marks the presence of the property and 0 the absence of the property.

This framework captures the nature of causation as *difference-making*: if variable *X* is causally relevant for variable *Y*, changes in the value of variable *X* make a difference in the value of variable *Y* (in a range of circumstances). Interventionist causation is also essentially *contrastive*: It is *X*'s taking some value *x* instead of *x'* that causes *Y*'s taking value *y* instead of *y'*.

The interventionist account accords well with the way causal notions are employed in the special sciences (Woodward, 2000, 2003, 2008). The account has also received broad acceptance among both philosophers and scientists. However, it seems to provide exactly the kind of "weak" notion of causation that Kim finds unsatisfactory. Kim is after a productive or generative notion of causation that is more metaphysically robust.

The main problem with such a stronger notion is that it would have to be somehow grounded in physics. In the end, the metaphysical question that Kim wants to answer is how there could be mental causes in a fundamentally physical world. If the stronger notion of causation was *not* grounded in physics, it is hard to see what reason there would be to prefer it to the interventionist account, assuming that the latter captures the notion of causation as it is needed in science and everyday life.

The problem with grounding causation in physics is that notions like cause and effect do not really play a role in our best physical theories (as famously argued by Bertrand Russell (1912-13), and more recently by Ladyman and Ross (2007), Loewer (2007), Norton (2007), and many others). The fundamental laws of physics relate the totality of a physical state at one time to the totality of the physical state at later instants, but do not single out causes and effects among these states. If we want to find causes that "bring about" or "produce" their effects, or causes that are "sufficient" for their effects, we have to consider something like the entire state of the universe as the cause for even a small effect.[9]

Of course, we can put labels onto relata that appear in physical equations and call some of them causes and others effects, but this is entirely superfluous to the physics itself. There is no "principle of causality" that would in any way guide or restrict physical

theory formation. Furthermore, there are cases even in Newtonian physics which go straight against our ideas of causation – for instance, effects that take place with no observable causes (Norton, 2007) – not to even speak of phenomena like quantum entanglement.

The interventionist account seems to capture the nature of causation both in special sciences and everyday life very well, and in fundamental physics, causal notions are unnecessary and superfluous.[10] It then seems that the interventionist account, insofar as it is successful, gives us all we want from an account of causation. A philosopher of mind could still insist that the question of what causation *really* is has to be answered. But from a scientific point of view, this search for the true nature of causation can be seen as just a metaphysical exercise. As Woodward (2008, 249) puts it: "We are thus left with possibility that the only people who think that vindicating the claim that mental states are causes requires showing that they are causes in a richer, more metaphysical sense are certain philosophers of mind."

3.3.2. Causal Exclusion in the Interventionist Framework

What are the consequences of the interventionist account for mental causation? Prima facie, it seems that mental causation is unproblematic in the interventionist framework. There are invariant psychological generalizations such that we can make interventions to mental states in order to change other mental states or physical behavior. For example, as Woodward (2008) points out, when you persuade someone, you manipulate her beliefs by providing information or material things, in order to change her other beliefs. Also many psychological and social science experiments involve intervening on the beliefs of the subjects, usually through verbal instruction, in order to change some other beliefs and observable behavior.

Upon closer philosophical analysis it appears that the interventionist account indeed vindicates mental causation. Several authors (e.g., Menzies, 2008, Raatikainen, forthcoming) have recently come up with an argument that claims to show that if the interventionist account is correct, mental states can be causes of physical behavior, and

they are not excluded by their physical realizers. This is due to the fact that causation in the interventionist account is a matter of difference-making, and not a matter of physically producing or bringing about the effect. The difference-making cause of a physical event, like a hand movement, can be a mental cause, and it is not excluded by some physical cause. Therefore, the exclusion principle does not hold or turns out to be nonsensical in the interventionist framework. On the other hand, Michael Baumgartner (forthcoming) and Vera Hoffmann-Kolss (unpublished manuscript) have argued that there is an interventionist version of the exclusion argument, and thus adopting the interventionist account does not make the problem of exclusion go away.

Instead of going through the details of these arguments, I will argue that there is a deeper underlying problem that applies to the arguments of participants at both sides of the debate. The problem is that typical causal representations of the mental and the physical causes fail to satisfy the *Causal Markov condition*.[11]

According to one formulation that is the most fitting one in the present context, the Causal Markov condition states (CM): conditional on its direct causes, each variable is independent of every other variable except its effects.[12] In other words, variables that are not related as cause or effect or as effects of a common cause have to be uncorrelated. It is widely agreed that when the causal relationships in a system are correctly and fully represented, CM will be satisfied. Furthermore, the condition *follows* from the interventionist definition of causation and some other plausible assumptions[13] (Hausman and Woodward, 1999). Hence, in a full and correct interventionist causal representation of a system, CM has to be satisfied.

Typical representations of mental causation in philosophy of mind, including the one applied in Kim's exclusion argument (section 2), *fail* to satisfy CM (see Figure 1). In these representations, mental property *M* causes another mental property *M**, physical (or neural) property *P* causes another physical (or neural) property *P**, *M* supervenes on *P*, and *M** supervenes on *P**. Due to supervenience, the values of *M* and *P* are correlated, and *M* depends on *P*. Whenever *M* changes, *P* also changes, and when the value of *P* is fixed, the value of *M* is also fixed.[14] However, *M* does not cause *P*, *P* does not cause *M*, and they are not both effects of a common cause. Hence, CM is violated. This means that something has gone wrong in building the causal representation of the system.

INSERT FIGURE 1 HERE

There are (at least) the following three ways of reacting to this problem complex. (1) The reductive solution: get rid of the mental variables, either by identifying them with physical variables or simply eliminating them. (2) The nonreductive solution: fix the level of analysis when building the causal representation, and never include supervenient variables in the same representation with their supervenient base variables. (3) Argue that this is a problem for the interventionist account, and that it needs to be replaced or revised (e.g., by adding some additional principles for dealing with supervenient variables).

The problem with the reductive solution is that if we accept it, we can just as well apply the same reasoning to nonmental variables, which leads to undesirable consequences. All that is required for the argument to work is that there is a supervenience relation between the variables, and supervenience relations can be found all over the place, also in biological, chemical, and even physical contexts. We can also consider the fact that the neural properties (variables) supervene on biochemical or some other lower-level properties (variables). Therefore, we can simply draw the same picture again, replacing mental variables by neural variables and neural variables by biochemical variables. Then it seems that since we got rid of the mental variables in the first case, we also have to get rid of the neural variables in the second case. Causation seems to be draining away towards some fundamental physical level, which is particularly strange if we consider the fact that there seems to be no causation at the fundamental physical level (see previous section).

This is a version of the *generalization argument* that has often been raised against Kim's exclusion argument (e.g., Block, 2003, van Gulick, 1992). The generalization argument states that if Kim's reasoning about mental properties is correct, then we can apply it to all higher-level or nonfundamental properties, which then are excluded. However, this is an absurd conclusion, so there has to be something wrong with Kim's argument. Kim has provided several answers to the generalization argument, but it is widely agreed that none of them is satisfactory (see, e.g., Walter, 2008). What a defender

of the exclusion argument (also the interventionist version) would have to show is that there is some principled reason why mental properties (variables) are excluded but other higher-level or macroproperties are not. Until such a reason is provided, the reductive solution is a nonstarter.

The *nonreductive* solution would be to allow higher-level causal representations, but not allow including the supervenient base variables in the same representation. For example, we would not include neural variables in the same representation as the mental variables. We would have a plurality of causal representation, but not representations that include both supervenient variables and their base variables. As Hausman and Woodward (1999, 531) put it in a different context: "One needs the right variables or the right level of analysis – variables that are sufficiently informative and that are not conceptually connected."

This solution is attractive and close to scientific practice, and I think ultimately something like this approach is the right way to go.[15] However, there are at least two problems. First of all, there seems to be an element of arbitrariness or *ad hoc* here, since the only reason for not including the supervenience base variables is that it would violate the Causal Markov condition. Secondly, there might be cases where we would like to include supervenient variables and their base variables in the same representation. If it turns out there are serious and scientifically relevant cases like that, it means trouble for the nonreductive solution.

Defending the nonreductive solution also requires showing what exactly goes wrong in the exclusion argument. At least one of the principles appealed to in the argument has to turn out false. Due to constraints of space, I cannot go into the details here, but the most likely candidate is the exclusion principle, which becomes highly problematic when it is formulated in interventionist terms (see Menzies (2008) and Raatikainen (forthcoming) for more). This is again due to the fact that interventionist causation is a matter of difference-making, not of physically producing the effect.

The third solution is to argue that the interventionist account of causation is deficient, and that we need to replace it, or at least revise it, for example by adding some further rules or principles for dealing with cases of supervenience. Baumgartner (forthcoming) argues that the exclusion argument is indeed a fundamental problem for

the interventionist account, and is skeptical regarding possible revisions. However, one argument against this solution is that if the problem arises only in an abstract philosophical context, like the problem of mental causation in philosophy of mind, it might be that the abstract philosophical context needs to be revised, not the interventionist account, which takes us back to options (1) and (2). Again, it remains to be seen how common or relevant are the cases where we want to include also supervenience base variables in the representation.

To summarize, Kim's argument does hold in one sense even in the interventionist framework: it shows that causal claims become very problematic when conjoined with supervenience claims. However, if it is understood as an argument to the effect that mental causation is not possible, or is more problematic than other kinds of causation, it does not hold.

Thus, with a correct understanding of causation, a large part of the motivation behind functional reduction disappears. Kim wanted to show that mental properties are functionally reducible in order to save mental causation. However, it seems that mental causation does not need such a rescue operation: mental causation in the interventionist sense is no more problematic than any other kinds of causation, and the search for more metaphysical (productive, generative, sufficient, etc.) mental causes is pointless.

What is then the motivation for reducing or reductively explaining the mental? I think the correct answer is that we want to reductively explain the mental because we want to explain everything there is to explain, and some kind of reductive explanation seems to be very fruitful in this context, as the success of neuroscience in recent decades shows. But what exactly is the nature of this explanatory enterprise?

4. Functional reduction as mechanistic explanation

Perhaps the functional model could be revised, taking into account all that has been said above, in roughly the following way. We want to reduce mental property *M*. First, we have to find out what the functional role of *M* is. However, this is not done through conceptual analysis alone, but through the interplay of conceptual analysis and empirical

research. Also, it is an ongoing process, and the initial definitions may be refined later. This first step is not necessarily temporarily prior to the next steps, and anyway the whole process is integrated and all the steps are intertwined. In the second step, we figure out what the neural mechanism that is the "realizer" of *M* is. *M* is neither identical to its realizer nor "just a concept" – the realizing mechanism is structured into levels, and *M* can be seen as a higher-level property of the mechanism. Third, we construct the "theory" that explains why the mechanism is the realizer of *M* – that is, we show how the functioning of the mechanism results in *M* (i.e., how the mechanism performs the functional role of *M*).

This quickly sketched revised account of functional reduction looks very much like what in philosophy of science is known as *mechanistic explanation*. The key idea of the mechanistic explanation paradigm (Bechtel, 2008, Bechtel and Richardson, 1993, Craver, 2007, Machamer et al., 2000) is that if one takes into account actual scientific practice in neuroscience and many of the life sciences, it turns out that instead of focusing on laws or formalizable theories, practicing scientists formulate explanations in terms of mechanisms.

According to an often-cited definition, mechanisms are to be understood as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, 3).). Or, as Bechtel (2008, 13) puts it, a "mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization." A mechanistic explanation then describes how the orchestrated functioning of the mechanism is responsible for the phenomenon to be explained.

This suggests that to functionally reduce a property *M* amounts to providing a mechanistic explanation for *M*. The upshot is that if we want to keep the model of functional reduction close to science, it turns out that there is no functional reduction over and above mechanistic explanation.

What does replacing the functional model with mechanistic explanation mean for the questions of reduction and causation? The mechanistic explanation model, conjoined with the interventionist account of causation, does not involve the kind of strong ontological reduction in terms of property identities or eliminations that Kim is after,

since it emphasizes the *multilevel* nature of mechanisms, and the causal and explanatory relevance of higher-level things. However, it is important to remember that the main reason for being an ontological reductionist (at least for Kim) is the causal exclusion argument. If the exclusion problem does not arise when we understand causation in interventionist terms, then also the motivation for being a strong reductionist fades away.

Many philosophers (e.g., Bechtel, 2008, Sarkar, 1992, Wimsatt, 1976) have argued that the process of "looking downward" and invoking parts of the mechanism to understand the behavior of the mechanism as a whole is close enough to what scientists generally take to be a reductive explanation to warrant treating the downward-looking aspect of mechanistic explanation as a kind of reductive explanation. On the other hand, Craver (2007) considers the framework of mechanistic explanation antireductive. This issue is mainly a terminological one, but I see no harm done calling downward-looking mechanistic explanation reductive explanation, as long as it is clearly distinguished from stronger forms of reduction. Regardless of whether we want to call mechanistic explanation reductive explanation, this approach supports a kind of causal and explanatory pluralism: higher-level entities or properties (including psychological entities and properties) do have causal and explanatory relevance, and are not reducible in any strong sense to lower-level entities and properties.

To conclude, functional reduction fails as a general account of reduction in philosophy of mind: if we try to understand it in a scientifically credible way, it effectively gives way to mechanistic explanation, which in turn leads to causal and explanatory pluralism. Whether this is compatible with "physicalism, or something near enough" (Kim, 2005) is an open question that has still to be addressed.

Acknowledgements

I am very grateful to Dan Brooks, Vera Hoffmann-Kolss, Jani Raerinne, Max Seeger, Achim Stephan and an anonymous referee for comments on earlier versions of this article. The article is based on a presentation at the workshop "Reductionism,

Explanation and Metaphors in the Philosophy of Mind" in September 2009 in Bremen, organized by Albert Newen and Raphael van Riel.

Notes

1. As an anonymous referee pointed out, one could argue that Nagelian reduction involves only the deduction of *laws*, which does not as such require formal theories. However, this only leads to a parallel problem: laws in the sense of generalizations that fill the traditional criteria for laws are not central in psychological and neuroscientific theories and explanations (Craver, 2007, Cummins, 2000, Machamer et al., 2000, Woodward, 2000).
2. Marras (2002, 2005), however, argues that functional reduction in fact collapses back to Nagelian reduction. I return to this in section 3.2.
3. Esfeld and Sachse (2007) have argued that by introducing functional sub-types we can have property identities and conservative functional reductions, multiple realizability notwithstanding.
4. This problem is obviously related to the issue of common-sense (analytical) vs. empirical functionalism (psychofunctionalism).
5. Perhaps one solution would be to argue that mental properties are some special kind of "abstract" properties. However, Kim does not appear to seriously consider such a solution. In any case, it would require developing or spelling out the metaphysics for such properties, which is no easy task.
6. In fact, Kim sometimes seems ready to reject the multiple realizability of mental properties and argues for "species-specific identities," such that "multiply realized properties are sundered into diverse realizers in different species and structures" (Kim, 1998, 105). This leads to problems if there is also multiple realizability within species or structures: it seems to follow that mental properties are spliced into properties restricted to very specific neural or physical structures, and it is hard to see how such properties could be relevant in scientifically explaining human behavior.
7. Wilson and Craver (2007) have recently defended a mechanistic approach to realization. They argue that this also comes close to how the term "realization" is often used in the cognitive sciences and neurosciences: when scientists say they are looking for the neural realization of memory consolidation, what they typically mean is that they are looking for the neural mechanism of memory consolidation. The approach of Wilson and Craver is promising, but remains rather provisional and schematic.
8. In section 4 I argue that we should understand reductive explanation as downward-looking mechanistic explanation. If there are one-to-many mappings from psychological properties or functions to the underlying mechanisms, this is no obstacle to downward-looking mechanistic explanation of those properties or functions. In these cases, different mechanisms can perform the same roughly defined function, and therefore there are different mechanistic explanations for this function. There is nothing problematic about this.
9. Perhaps it is sufficient to consider the state of the universe on the surface of a sphere with a radius of about 300 000 000 meters centered on the effect, assuming that the cause

precedes the effect by one second – the speed of causal influence cannot be faster than the speed of light. Of course, this does not make the idea of productive physical causation any less problematic. See Loewer (2007) for more.

10. As an anonymous referee pointed out, not all philosophers of physics agree that there is no causation in fundamental physics (see, e.g., Frisch, 2009). However, even if it turns out that causal notions do play a role in fundamental physics, it is still the case that there is currently no metaphysically robust and physically grounded notion of causation that would be suitable for considering mental causation and a serious alternative to interventionist causation.

11. This was pointed out to me by Dan Brooks, for which I am very grateful.

12. See Hausman and Woodward (1999) for other formulations and an extensive discussion of the Causal Markov condition. Another condition that is also extensively covered in the same paper, and that could perhaps also be used as a basis for the arguments in this section, is modularity: a system consisting of several causal relationships is modular to the extent that these various causal relationships can be changed or disrupted while leaving the others intact. Both CM and modularity have been under intense discussion in recent years – see, for example, Cartwright (2002) or Steel (2006).

13. Alternatively, it could be said that the interventionist definition follows from CM and some other plausible assumptions. Without (something like) CM it is impossible to talk of variables being causal in the interventionist sense.

14. According to a standard definition, a set of A-properties supervenes on a set of B-properties if and only if two things cannot differ with respect to their A-properties without also differing with respect to the B-properties.

15. Recently Shapiro and Sober (2007) have also argued that supervenient causes are problematic in the interventionist framework and defended a nonreductive solution. Let us consider a situation where we want examine whether M, which supervenes on P, is a cause of physical behavior B. We have to make an intervention on M such that other causes of B, including P, remain unchanged. The problem is that this is impossible, since the value of P determines the value of M (due to supervenience). It is not acceptable or nomologically possible to wiggle M while holding P fixed. Hence, this must be a wrong way of conceptualizing the situation.

References

Baumgartner, Michael, forthcoming: Interventionism and epiphenomenalism. In: *Canadian Journal of Philosophy*.

Bechtel, William, 2008: *Mental Mechanisms*. London: Routledge.

Bechtel, William, and Robert C. Richardson, 1993: *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Bickle, John, 1998: *Psychoneural Reduction: The New Wave*. Cambridge, MA: MIT Press.

Bickle, John, 2008: Real Reduction in Real Neuroscience: Metascience, Not Philosophy of Science (and Certainly Not Metaphysics!). In: Hohwy, J.; Kallestrup, J. (eds.) *Being Reduced*. Oxford: Oxford University Press, pp. 34-51.

Bickle, John, forthcoming: The Changing Faces and Scientific Bases of Mind-Brain Reductionism. In: *Reti, saperi, linguaggi* (Journal of the Department of Cognitive Science, University of Messina, Italy).

Block, Ned, 2003: Do Causal Powers Drain Away? In: *Philosophy and Phenomenological Research* 67, pp. 133-150.

Cartwright, Nancy, 2002: Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward. In: *British Journal for the Philosophy of Science* 53, pp. 411-453.

Chalmers, David J., 1996: *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, David J., and Frank Jackson, 2001: Conceptual Analysis and Reductive Explanation. In: *The Philosophical Review* 110, pp. 315-360.

Craver, Carl F., 2007: *Explaining the Brain*. Oxford: Oxford University Press.

Cummins, Robert 2000. "How Does it Work?" versus "What Are the Laws?" Two Conceptions of Psychological Explanation. In: Keil, F.; Wilson, R. (eds.): *Explanation and Cognition*. Cambridge: MIT Press, pp. 117-144.

Esfeld, Michael, and Christian Sachse, 2007: Theory Reduction by Means of Functional Sub-types. In: *International Studies in the Philosophy of Science* 21, pp. 1-17.

Frisch, Mathias, 2009: 'The Most Sacred Tenet'? Causal Reasoning in Physics. In: *British Journal for the Philosophy of Science* 60, pp. 459-474.

Hardcastle, Valerie, 2001: The Nature of Pain. In: Bechtel, W.; Mandik, P.; Mundale, J.; Stufflebeam, R.S. (eds.): *Philosophy and the Neurosciences: A Reader*. Malden, MA: Blackwell, pp. 295-311.

Hausman, Daniel M., and James Woodward, 1999: Independence, Invariance and the Causal Markov Condition. In: *British Journal for the Philosophy of Science* 50, pp. 521-583.

Hoffmann-Kolss, Vera (unpublished manuscript). The Supervenience Argument Is Alive and Kicking.

Hooker, Clifford A., 1981: Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction. In: *Dialogue* 20, pp. 38-59, 201-236, 496-529.

Khalidi, Muhammad A., 2005: Against Functional Reductionism in Cognitive Science. In: *International Studies in the Philosophy of Science* 19, pp. 319-333

Kim, Jaegwon, 1992: Multiple Realization and the Metaphysics of Reduction. In: *Philosophy and Phenomenological Research* 52, pp. 1–26.

Kim, Jaegwon, 1998: *Mind in a Physical World*. Cambridge, MA: MIT Press.

Kim, Jaegwon, 1999: Making Sense of Emergence. In: *Philosophical Studies* 95, pp. 3-36.

Kim, Jaegwon, 2002: Mental Causation and Consciousness: The Two Mind-body Problems for the Physicalist. In: Gillett, C; Loewer, B. (eds.): *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, pp. 271-283.

Kim, Jaegwon, 2005: *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kim, Jaegwon, 2008: Reduction and Reductive Explanation: Is One Possible without the Other? In: Hohwy, J.; Kallestrup, J. (eds.): *Being Reduced*. Oxford: Oxford University Press, pp. 93-114.

Kim, Jaegwon, 2009: Mental Causation. In: McLaughlin, B; Beckermann, A.; Walter, S. (eds.): *The Oxford handbook of philosophy of mind*. Oxford: Oxford University Press, pp. 29-52.

Kistler, Max, 2005: Is Functional Reduction Logical Reduction? In: *Croatian Journal of Philosophy* 14, pp. 219-234.

Ladyman, James, and Don Ross, 2007: *Every Thing Must Go: Metaphysics Naturalised*. Oxford: Oxford University Press.

Levine, Joseph, 1993: On Leaving Out What It's Like. In: Humphreys, G.; Davies, M. (eds.): *Consciousness*. Oxford: Blackwell, pp. 121–136.

Lewis, David, 1972: Psychophysical and Theoretical Identifications. In: *Australasian Journal of Philosophy* 50, pp. 249-258.

Loewer, Barry, 2007: Mental Causation, or Something Near Enough. In: McLaughlin, B.; Cohen, J. (eds.): *Contemporary Debates in Philosophy of Mind*. Malden, MA: Blackwell Publishing, pp. 243-264.

Machamer, Peter K., Lindley Darden, and Carl Craver, 2000: Thinking About Mechanisms. In: *Philosophy of Science* 67, pp. 1-25.

Marras, Ausonio, 2002: Kim on Reduction. In: *Erkenntnis* 57, pp. 231-257.

Marras, Ausonio, 2005: Consciousness and Reduction. In: *British Journal for the Philosophy of Science* 56, pp. 335-361.

McCauley, Robert N., 2007: Reduction: Models of Cross-scientific Relations and their Implications for the Psychology-Neuroscience Interface. In Thagard, P. (ed.): *Handbook of the Philosophy of Psychology and Cognitive Science*. Amsterdam: Elsevier, pp. 105-158.

Menzies, Peter, 2008: The Exclusion Problem, the Determination Relation, and Contrastive Causation. In Hohwy, J.; Kallestrup, J. (eds.): *Being Reduced*. Oxford: Oxford University Press, pp. 196-217.

Menzies, Peter, and Huw Price, 1993: Causation as a Secondary Quality. In: *The British Journal for the Philosophy of Science* 44, pp. 187-203.

Nagel, Ernest, 1961: *The Structure of Science*. London: Routledge & Kegan Paul.

Norton, John D., 2007: Causation as Folk Science. In: Price, H.; Corry, R. (eds.) *Causation, Physics, and the Constitution of Reality. Russell's Republic Revisited*. Oxford: Oxford University Press, pp. 11-44.

Pearl, Judea, 2000: *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Polger, Thomas W., 2004: *Natural Minds*. Cambridge: MIT Press.

Polger, Thomas W., 2007: Realization and the Metaphysics of Mind. *Australasian Journal of Philosophy* 85, pp. 233-259.

Polger, Thomas, and Lawrence Shapiro, 2008: Understanding the Dimensions of Realization. In: *The Journal of Philosophy* 105, pp. 213-222.

Putnam, Hilary, 1967: Psychological Predicates. In: Capitan, W.H.; Merrill, D.D. (eds.): *Art, Mind, and Religion*. Pittsburg: Pittsburg University Press, pp. 37–48.

Raatikainen, Panu, forthcoming: Causation, Exclusion, and the Special Sciences. In: *Erkenntnis*.

Russell, Bertrand, 1912-1913: On the Notion of Cause. In: *Proceedings of the Aristotelian Society* 13, pp. 1-26.

Sarkar, Sahotra, 1992: Models of Reduction and Categories of Reductionism. In: *Synthese* 91, pp. 167–194.

Shapiro, Lawrence A., 2004: *The Mind Incarnate*. Cambridge, MA: MIT Press.

Shapiro, Lawrence A., and Elliott Sober, 2007: Epiphenomenalism – the Do's and the Don'ts. In: Wolters, G.; Machamer, P. (eds.): *Thinking about Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press, pp. 235-264.

Spirtes, Peter, Clark Glymour, and Richard Scheines, 1993: *Causation, Prediction, and Search*. New York: Springer.

Steel, Daniel, 2006: Comment on Hausman & Woodward on the Causal Markov Condition. In: *British Journal for the Philosophy of Science* 57, pp. 219-231.

van Gulick, Robert, 1992: Three Bad Arguments for Intentional Property Epiphenomenalism. In: *Erkenntnis* 36, pp. 311-332.

Walter, Sven, 2008: The Supervenience Argument, Overdetermination, and Causal Drainage: Assessing Kim's Master Argument. In: *Philosophical Psychology* 21, pp. 671–694.

Walter, Sven, and Markus I. Eronen (forthcoming): Reductionism, Multiple Realizability, and Levels of Reality. In: French, S.; Saatsi, J. (eds.): *Continuum Companion to the Philosophy of Science*. Continuum.

Wilson, Robert A., and Carl F. Craver, 2007: Realization: Metaphysical and Scientific Perspectives. In: Thagard, P. (ed.): *Handbook of the Philosophy of Psychology and Cognitive Science*. Amsterdam: Elsevier, pp. 81-104

Wimsatt, William C., 1976: Reductionism, Levels of Organization, and the Mind-Body Problem. In: Globus et al. (eds.): *Consciousness and the Brain. A Scientific and Philosophical Inquiry*. New York: Plenum Press, pp. 205-267.

Woodward, James, 2000: Explanation and Invariance in the Special Sciences. In: *The British Journal for the Philosophy of Science* 51, pp. 197-254.

Woodward, James, 2003: *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, James, 2008: Mental Causation and Neural Mechanisms. In Hohwy, J.; Kallestrup, J. (eds.): *Being Reduced*. Oxford: Oxford University Press, pp. 218–262.

Woodward, James, and Christopher Hitchcock, 2003: Explanatory Generalizations, Part I: A Counterfactual Account. In: *Noûs* 37, pp. 1-24.